Influence of binary mask estimation errors on robust speaker identification

May, Tobias

Published in:
Speech Communication

Link to article, DOI:
10.1016/j.specom.2016.12.002

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

Speech Communication 87 (2017) 40–48

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Influence of binary mask estimation errors on robust speaker identification

Tobias May*

Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Kgs. Lyngby DK - 2800, Denmark

ARTICLE INFO

Article history:
Received 8 March 2016
Revised 5 December 2016
Accepted 9 December 2016
Available online 16 December 2016

Keywords:
Speaker identification
Missing data
Ideal binary mask
Estimated binary mask
Bounded marginalization
Full marginalization
Direct masking

ABSTRACT

Missing-data strategies have been developed to improve the noise-robustness of automatic speech recognition systems in adverse acoustic conditions. This is achieved by classifying time-frequency (T-F) units into reliable and unreliable components, as indicated by a so-called binary mask. Different approaches have been proposed to handle unreliable feature components, each with distinct advantages. The direct masking (DM) approach attenuates unreliable T-F units in the spectral domain, which allows the extraction of conventionally used mel-frequency cepstral coefficients (MFCCs). Instead of attenuating unreliable components in the feature extraction front-end, full marginalization (FM) discards unreliable feature components in the classification back-end. Finally, bounded marginalization (BM) can be used to combine the evidence from both reliable and unreliable feature components during classification. Since each of these approaches utilizes the knowledge about reliable and unreliable feature components in a different way, they will respond differently to estimation errors in the binary mask. The goal of this study was to identify the most effective strategy to exploit knowledge about reliable and unreliable feature components in the context of automatic speaker identification (SID). A systematic evaluation under ideal and non-ideal conditions demonstrated that the robustness to errors in the binary mask varied substantially across the different missing-data strategies. Moreover, full and bounded marginalization showed complementary performances in stationary and non-stationary background noises and were subsequently combined using a simple score fusion. This approach consistently outperformed individual SID systems in all considered experimental conditions.

1. Introduction

The automatic identification of speakers in adverse scenarios is an important building block for many applications, including access control, authentication, personalization of communication services and forensic applications (Campbell, 1997; Campbell et al., 2009). Conventional speaker identification (SID) systems employ mel-frequency cepstral coefficients (MFCCs) in combination with Gaussian mixture model (GMM) classifiers (Reynolds and Rose, 1995) and universal background model (UBM) adaptation (Reynolds et al., 2000). To increase the robustness of SID systems against changes in the acoustic environment between the training and testing stage, feature normalization strategies are usually applied. The most commonly used normalization strategies are file-based or adaptive cepstral mean and variance normalization (Viikki and Laurila, 1998) as well as histogram equalization (HEQ) techniques (de la Torre et al., 2005).

In contrast to conventional approaches, the missing-data technique classifies the time-frequency (T-F) representation of noisy speech into target-dominated (reliable) and interference-dominated (unreliable) T-F units (Cooke et al., 2001). This binary decision can, for example, be accomplished by the so-called ideal binary mask (IBM), which assumes *a priori* knowledge about the energy of the target and the interfering noise (Wang, 2005). The recognition of speaker identities is subsequently performed using only those T-F units that are believed to be reliable. Many studies have demonstrated that the missing-data technique improves the robustness of automatic speech recognition (Vizinho et al., 1999; Cooke et al., 2001; Ris and Dupont, 2001) and SID systems (Drygajlo and El-Maliki, 1998; Jančovič and Köküer, 2006; Shao and Wang, 2006; May et al., 2012a; 2012b; Zhao et al., 2014) in noisy environments. If, instead of a binary classification, an estimation of the feature uncertainty is available, this information can be ex-

* Corresponding author.
E-mail address: tobmay@elektro.dtu.dk

http://dx.doi.org/10.1016/j.specom.2016.12.002

ploited by an uncertainty decoder for improved robustness (Deng et al., 2005; Shao et al., 2007; Ozerov et al., 2013; Yu et al., 2014). For a comprehensive overview the reader is referred to Kolossa and Haeb-Umbach (2011).

Assuming that a binary decision about the feature reliability is available, two different approaches exist to incorporate the knowledge about reliable and unreliable T-F units in the classification back-end. The first - full marginalization (FM) - ignores all unreliable T-F units and estimates the likelihood of a particular speaker identity based on reliable T-F units only. Despite being dominated by noise, unreliable T-F units can be used to provide *counter-evidence* based on the concept of energetic masking (Cooke et al., 2001). Consequently, the second approach - bounded marginalization (BM) - additionally exploits knowledge about unreliable T-F units for improved identification performance by effectively penalizing less energetic speaker models.

One of the most critical factors that limits the performance of missing-data strategies is the accuracy of the binary mask. When *a priori* knowledge in form of the IBM is available, the BM approach is very effective and produces SID scores above 95%, even at negative signal-to-noise ratios (SNRs) (May et al., 2012b). However, when estimated binary masks (EBMs) are used instead, the achievable performance of BM decreases due to misclassified T-F units in the binary mask, most noticeably in highly non-stationary background noises (May et al., 2012b). In addition, BM treats reliable and unreliable feature components differently during classification and, thus, requires a correspondence between the individual T-F units in the binary mask and the feature components used for recognition. Therefore, the BM approach is limited to spectral features, such as filter-bank energy (FBE) features, which are known to be less powerful than their decorrelated counterparts, such as MFCC features, which operate in the cepstral domain.

The performance of FBE features can be improved by applying a derivative finite impulse response (FIR) filter across frequency channels, which produces decorrelated filter-bank energy (DFBE) features (Nadeu et al., 1995; 2001). These DFBE features achieve similar SID performance compared to MFCC features, with the advantage that frequency-specific distortions due to background noise remain local and can still be associated with a restricted set of T-F units. While it is not beneficial to exploit information about unreliable T-F units after the FIR-based decorrelation stage due to the wide feature bounds (Barker, 2012), DFBE features can be combined with FM, where unreliable feature components are ignored during classification.

Alternatively, the knowledge about reliable and unreliable T-F units can be exploited in the feature extraction front-end. Specifically, the spectral representation of noisy speech can be enhanced by a binary or continuous gain function (El-Solh et al., 2007; Sadjadi and Hansen, 2010; Jensen and Hendriks, 2012; Godin et al., 2013). This allows the use of conventional decorrelation stages, such as the discrete cosine transform (DCT), to convert the modified FBE features to MFCC features. When assuming *a priori* knowledge about the noise power, it was shown that an *ideal* noise reduction scheme based on a continuous gain function achieved the same SID performance compared to IBM-based BM (May et al., 2012b). Likewise, the direct masking (DM) approach applies the IBM directly to the FBE features, and was reported to produce similar automatic speech recognition performance compared to the BM strategy (Hartmann et al., 2013). However, estimation errors in the gain function can distort the resulting feature vector, which may limit the effectiveness under realistic conditions.

Each of the aforementioned missing-data methods has distinct advantages and applies the binary mask in a different way. As a result, mask estimation errors will have different consequences on the achievable speaker identification performance. Nevertheless, the influence of binary mask errors on the different missing-data strategies has not yet been systematically investigated. The majority of studies obtained an estimation of the binary mask by predicting the local SNR in individual T-F units (Drygajlo and El-Maliki, 1998; Renevey and Drygajlo, 2000; Cooke et al., 2001; Ris and Dupont, 2001; May et al., 2012b). Recently, supervised learning approaches have reported increased mask estimation accuracies by exploiting *a prior* knowledge about the distribution of acoustic features observed during an initial training phase (Seltzer et al., 2004; May and Dau, 2013; 2014; Zhao et al., 2014). However, the general advantage of missing-data strategies is that speaker models are trained with clean speech only, and no prior knowledge about the acoustic environment or about the interfering noise is required.

SID systems based on the GMM-UBM back-end are the most commonly-used systems when applying missing-data strategies (Togneri and Pullella, 2011; May et al., 2012a; 2012b; Zhao et al., 2012; 2014). Lately, the i-vector approach has received increasing attention, in particular in the field of speaker verification, by considering the inter- and intra-speaker variability of the feature vector (Dehak et al., 2011). A recent comparative study of noise reduction strategies (e.g. power spectral subtraction, Wiener filtering and log-minimum mean-square error speech enhancement) reported similar relative improvements over a MFCC baseline system for both GMM-UBM and i-vector-based SID systems (Godin et al., 2013). From this perspective, a similar benefit can be expected when the DM approach is combined with i-vector-based SID systems.

The goal of this study was to determine the effectiveness of full marginalization, bounded marginalization and direct masking in the context of closed-set SID under ideal and non-ideal conditions. To facilitate a proper comparison, a unified framework was used to optimize the criteria for deriving ideal and estimated binary masks for each missing-data system separately. In non-ideal conditions, EBMs were obtained by applying a threshold criterion to the estimated speech presence probability (SPP) in individual T-F units, which was shown to produce competitive results compared to supervised learning approaches (May and Gerkmann, 2014). A systematic comparison between ideal and non-ideal conditions was performed to analyze the robustness of the different missing-data approaches to estimation errors in the binary mask. Moreover, the SID performance was compared to a conventional noise reduction scheme. Afterwards, the noise-robustness of the most successful missing-data systems was compared to a conventional MFCC system across a wide range of acoustic conditions. Finally, a simple score fusion was tested which combined two missing-data systems with complementary advantages in stationary and non-stationary background noises. To support reproducible research, the complete Matlab code of the SID framework, which was used to produce all experimental results in the present study, is available online (May, 2016).

2. System

The SID framework shown in Fig. 1 consisted of a training and a testing stage. First, speaker models were trained with features extracted from clean speech. In the testing stage, features were derived from noisy speech and the reliability of individual feature components was estimated by means of a binary mask. The subsequent classification stage was configured to use three different missing-data strategies, namely full marginalization (FM), bounded marginalization (BM) and direct masking (DM). Depending on which type of missing-data strategy was used, different feature decorrelation and normalization stages were combined. A list of tested configurations is shown in Table 1. Each processing stage is described in detail in the following.

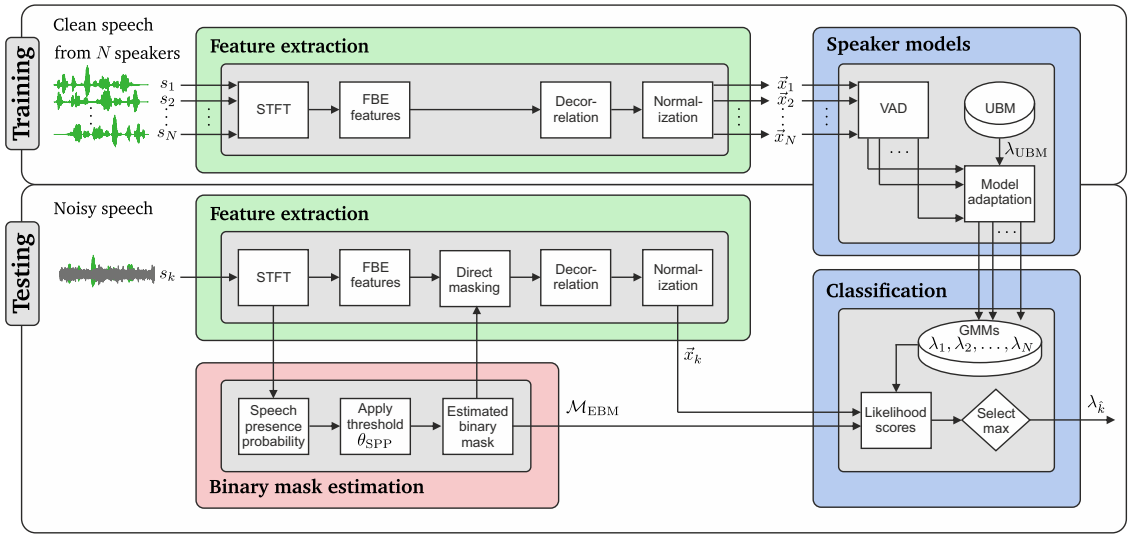

Fig. 1. Block diagram showing the training and the testing stage of the speaker identification framework.

Table 1
Configurations of evaluated SID systems.

Method	Features	Direct masking	Decor-relation	Normal-ization	Classifier
FBE BM	FBE	no	no	no	BM
DFBE FM	FBE	no	FIR	HEQ	FM
DFBE DM	FBE	yes	FIR	HEQ	GMM
MFCC DM	FBE	yes	DCT	HEQ	GMM

2.1. Filter-bank energy features

The input signal was sampled at a rate of 16 kHz and divided into overlapping frames of 20 ms duration with a shift of 10 ms. Each frame was Hamming windowed and zero-padded to a length of 512 samples and a short-time discrete Fourier transform (STFT) was computed. The STFT magnitudes were multiplied by 32 auditory filters whose center frequencies were equally spaced on the mel-frequency scale between 80 and 8000 Hz, resulting in an auditory spectrogram. This auditory spectrogram was cube-root compressed to produce the final set of FBE features used for identification. Cube-root compression is often employed in missing-data systems because it allows for a proper definition of the lower feature bound, which is commonly set to zero.

2.2. Binary masks

2.2.1. Ideal binary mask

The IBM requires *a priori* information about the target and the masker. Assuming knowledge about the energy of both the target and masker in individual T-F units, the *true* local SNR was computed. Subsequently, the IBM $\mathcal{M}_{\text{IBM}}(t, f)$ was determined by comparing the SNR in time frame t and frequency channel f to a predefined local criterion (LC)

$$\mathcal{M}_{\text{IBM}}(t, f) = \begin{cases} 1 & \text{if } SNR(t, f) > \text{LC} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

2.2.2. Estimated binary mask

In practical applications, the IBM is not available and, hence, has to be blindly estimated from the noisy speech signal. In this study, the binary mask was estimated by an algorithm that does not require any *a priori* knowledge about the target or the interfering signal. More specifically, the EBM was derived from noisy speech by first estimating the *a posteriori* SPP in the STFT domain using Gerkmann and Hendriks (2012). Here, the same STFT parameters as described in Section 2.1 were used. This discrete Fourier transform domain SPP was integrated into 32 auditory filters and averaged across a spectro-temporal neighborhood function, which substantially improved the accuracy of the EBM (May and Gerkmann, 2014). Following May and Gerkmann (2014), a plus-shaped neighborhood function spanning over 5 time frames and 5 auditory filters was applied. Finally, the estimated binary mask $\mathcal{M}_{\text{EBM}}(t, f)$ was obtained by comparing the SPP in the auditory domain, denoted as $\widetilde{\mathcal{P}}(t, f)$, in time frame t and frequency channel f to a threshold

$$\mathcal{M}_{\text{EBM}}(t, f) = \begin{cases} 1 & \text{if } \widetilde{\mathcal{P}}(t, f) > \theta_{\text{SPP}} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

2.3. Direct masking

The binary mask was applied as a binary gain function directly to the compressed FBE features to attenuate noise-dominated feature components. The amount of noise reduction was limited by a lower floor β in order to reduce the impact of distortions (musical noise) caused by the binary processing. Commonly used floor values are within the range of 20–26 dB (Berouti et al., 1979; Anzalone et al., 2006; Zhao et al., 2014). In this study, β was set to 26 dB according to Zhao et al. (2014).

2.4. Decorrelation stage

2.4.1. FIR-based frequency filtering

A second-order FIR was used to filter the compressed FBE features across frequency to produce decorrelated filter-bank energy (DFBE) features (Nadeu et al., 2001). In contrast to the conventionally-used DCT, the FIR filtering maintains the frequency axis of the feature space by combining information present in two neighboring frequency channels. As a result, the influence of interfering noise in a restricted frequency region can still be assigned to a limited frequency range in the DFBE feature space, which makes it possible to use DFBE features in combination with full marginalization.

2.4.2. DCT

The classical MFCC features were obtained by applying a DCT to the FBE features (Davis and Mermelstein, 1980). The first basis

function of the DCT is related to the overall energy and, therefore, is strongly affected by interfering noise. Thus, the first MFCC feature was omitted for improved noise-robustness and 13 coefficients (2nd–14th) were retained.

2.5. Feature normalization

To improve the ability of the trained speaker models to deal with acoustic conditions that differed from the clean training data, the extracted features were normalized by file-based HEQ (de la Torre et al., 2005). Instead of normalizing only the feature mean and the variance (Viikki and Laurila, 1998), HEQ equalizes all moments of the feature distribution. The HEQ was performed with a mapping function consisting of 100 percentiles using the reference implementation from Schädler and Kollmeier (2015).

2.6. Classification

Each classifier produced a matrix of likelihood scores, which contained the frame-based likelihoods for each of the enrolled speaker identity. Equal prior probabilities were assumed for all speakers.

2.6.1. Gaussian mixture models

A conventional GMM classifier was used to calculate frame-based likelihoods (Reynolds and Rose, 1995).

2.6.2. Full marginalization

Given a binary mask, the classifier was modified to discard unreliable T-F units during identification and computed the frame-based likelihoods based on reliable T-F units only (Cooke et al., 2001).

2.6.3. Bounded marginalization

Given a binary mask, the BM classifier utilized knowledge about both reliable and unreliable T-F units (Cooke et al., 2001). The lower bound was set to zero and the upper bound for each unreliable T-F unit was set to the corresponding FBE feature value.

2.7. Score fusion and identification

A simple score fusion approach according to Zhao et al. (2012); 2014) was implemented to combine the evidence provided from multiple SID systems. First, a vector of SID scores s was produced for each system by adding the frame level log-likelihood scores across all frames. Then, this vector of SID scores s was normalized across all enrolled speaker identities to a range between [0, 1] for each individual SID system separately

$$\hat{s} = \frac{s - \min(s)}{\max(s) - \min(s)}, \tag{3}$$

and subsequently combined across SID systems on a sentence-by-sentence basis. The SID decision was obtained by selecting the speaker identity that maximized the final SID score.

3. Evaluation

3.1. Databases

Closed-set speaker identification experiments were conducted using two different databases, namely the EMIME database[1] (Wester, 2010) consisting of 56 speakers and the TIMIT database (Garofolo et al., 1993) containing 630 speakers. The smaller EMIME database was used as a validation set, whereas large-scale comparisons were performed using the TIMIT database. In general, a restricted subset of 10 sentences was randomly selected for each speaker, from which 8 sentences were used for training (see Section 3.2) and the remaining 2 sentences were mixed with noise at various SNRs and subsequently used for evaluation (see Section 3.3). To reduce the influence of this randomized selection, the speaker identification performance was averaged over a series of 10 simulations, each containing a new set of randomly selected sentences for training and testing.

The EMIME database consists of bilingual recordings from four different languages (English, Finnish, German and Mandarin). For each language, a set of 145 sentences is available for 14 speakers (7 male and 7 female talkers) in their respective mother tongue and in English. For the SID experiments reported in this study, the English recordings based on the close-talking microphone were used, resulting in a closed set of 56 speakers (28 male and 28 female talkers).

The TIMIT database contains 10 phonetically balanced sentences from 630 speakers (438 males and 192 females), forming a set of 6300 sentences (Garofolo et al., 1993). The two dialect sentences (the SA sentences) were the same across all 630 speakers, and, therefore, were always included in the training set.

3.2. Model training

All SID systems were trained with features extracted from clean speech. A simple energy-based voice activity detector was used during training to consider only features with relevant speech activity. Speech activity was detected if the frame-based energy was within 40 dB of the global maximum measured across each sentence (Kinnunen and Li, 2010).

Speaker models were trained using the following two-step procedure. First, a UBM was constructed by training a 128-component GMM with diagonal covariance matrices using the pooled speech material from all enrolled speakers. Second, speaker-specific GMM models were obtained by adapting the mean vectors of the UBM to the speaker-specific speech material using 5 iterations and a relevance factor of 16 (Reynolds et al., 2000). The UBM adaptation of speaker models has been shown to substantially improve the performance of missing-data SID systems, in particular for non-stationary background noises (May et al., 2012b). All GMMs were implemented using the NETLAB package (Nabney and Bishop, 2004). Whenever the TIMIT database was used, the UBM was trained with the remaining speech material from 430 speakers that were not used for the actual identification experiment in order to support sequential enrollment of speaker identities.

3.3. Evaluation

The speaker identification systems were tested with noisy speech. Therefore, the clean speech material was corrupted with different noise types at −5, 0, 5, 10 and 15 dB SNR. The following five background noises were used: two types of speech-shaped noise (stationary ICRA1-noise and non-stationary ICRA7-noise (Dreschler et al., 2001)), 0.5-Hz amplitude-modulated white noise, as well as factory noise and cockpit noise from the NOISEX database (Varga and Steeneken, 1993). The SID performance was evaluated by comparing the estimated identity with the real speaker identity on a sentence-by-sentence basis.

In addition, the accuracy of the binary decisions was quantified by comparing the similarity between the estimated binary mask and the ideal binary mask, which was computed with an LC of 0 dB. First, the hit rate (HIT; percentage of correctly classified speech-dominated T-F units) minus the false alarm rate (FA; percentage of erroneously identified noise dominated T-F units)

[1] The EMIME database is available at http://www.emime.org

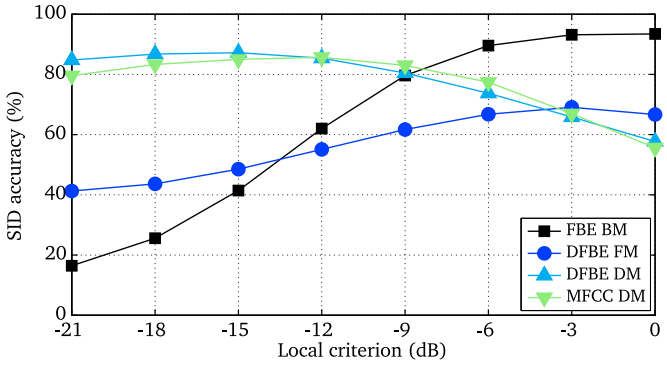

Fig. 2. Influence of the LC on IBM-based SID performance for different missing-data systems. The SID performance was evaluated for 56 speakers at 0 dB SNR and averaged across all five noise types.

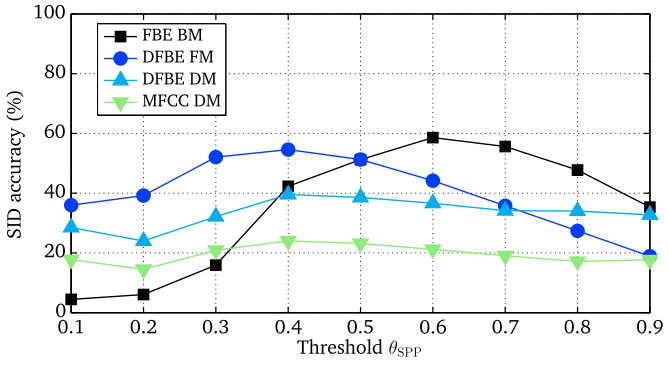

Fig. 3. Influence of the threshold parameter θ_{SPP} on EBM-based SID performance for different missing-data systems. The SID performance was evaluated for 56 speakers at 0 dB SNR and averaged across all five noise types.

was computed. The HIT - FA metric is often used to evaluate binary speech segregation systems (Kim et al., 2009). Furthermore, the overall percent of misclassified T-F units (accuracy) was measured, reflecting both the correctly classified target-dominated and the correctly classified noise-dominated T-F units. Finally, the mask density was determined, which indicates the percentage of reliable T-F units in the binary mask.

4. Experiments

A series of five closed-set SID experiments was conducted. The first two experiments evaluated the four missing-data systems (listed in Table 1) under ideal and non-ideal conditions with 56 speakers from the EMIME database. Specifically, the first experiment used IBMs (see Section 2.2.1) and analyzed the influence of the LC. The second experiment employed EBMs (see Section 2.2.2) and investigated the role of the threshold parameter θ_{SPP}. Based on these first two experiments, optimal threshold parameters were derived for all missing-data systems and kept constant across the remaining experimental conditions. The third experiment analyzed the sensitivity of the DM approach to errors in the binary mask by comparing the performance of EBMs with estimated ratio masks (ERMs) and a conventional noise reduction algorithm. The last two experiments used 200 speakers from the TIMIT database and evaluated the SID performance over a wide range of SNRs. The fourth experiment compared the best performing missing-data strategies to a conventional MFCC system. Finally, the fifth experiment combined two missing-data approaches with complementary advantages in stationary and non-stationary background noises.

4.1. Experiment 1: SID performance using IBMs

The SID performance of the four missing-data systems using IBMs is shown in Fig. 2 as a function of the LC. The SID experiment was conducted with 56 speakers from the EMIME database at 0 SNR and the results were averaged across 10 simulations and all five background noises. The highest SID performance of 93.1% was achieved by the bounded marginalization method ("FBE BM") with an LC of 0 dB. Because this system utilized both reliable and unreliable T-F units for the likelihood calculations, a balanced threshold around 0 dB was expected to be optimal. The two systems that combined the direct masking stage with either MFCC ("MFCC DM") or DFBE features ("DFBE DM") performed fairly similar, with optimal LCs at −12 dB and −18 dB leading to 85.8% and 87.4% identification performance, respectively. Compared to BM, the lower SID accuracy of the DM approach can be attributed to the fact that unreliable feature components were attenuated, thus ignoring potential speaker-specific differences. Full marginalization ("DFBE FM") discarded unreliable T-F units and only used reliable T-F units for identification. The obtained SID performance of 68.4% was substantially lower compared to the other methods, with an optimal LC of −3 dB.

4.2. Experiment 2: SID performance using EBMs

The SID performance for the missing-data systems employing the EBM is presented in Fig. 3 as a function of the threshold parameter θ_{SPP}. Similar to the previous experiment, a set of 56 speakers was tested at 0 dB SNR and results were averaged across 10 simulations and all five noise types. In addition, the quality of the EBM was evaluated by two technical measures, namely the HIT - FA and the accuracy (see Section 3.3), and both are shown along with the mask density in Table 2.

Bounded marginalization ("FBE BM") showed the highest SID performance of 59.1% with an optimal speech presence probability threshold of $\theta_{SPP} = 0.6$. Furthermore, the technical EBM evaluation presented in Table 2 revealed that this threshold produced a high binary mask accuracy of 85.7%, reflecting the requirement that both speech-dominated and noise-dominated T-F units in the EBM must be correctly estimated with similar priority. For all other missing-data methods, the highest SID accuracy was obtained with a SPP threshold of $\theta_{SPP} = 0.4$. Interestingly, this threshold coincided with the highest HIT - FA of 54.9%, as indicated in Table 2. Full marginalization ("DFBE FM") was almost as good as BM and achieved 54.9% SID accuracy. Although the FM strategy showed the lowest performance of all missing-data methods under ideal conditions in the previous experiment, it was apparently more robust against estimation errors in the EBM. In contrast, the SID systems based on direct masking were less effective when a realistic amount of speech-dominated T-F units was misclassified in the binary mask. In addition, there was a substantial performance

Table 2
Evaluation of the EBM as a function of the threshold parameter θ_{SPP}.

Threshold θ_{SPP}	Metric (%)		
	HIT - FA	Accuracy	Density
0.1	5.5	22.6	93.2
0.2	27.9	45.4	68.3
0.3	48.5	68.1	42.8
0.4	**54.9**	80.0	27.9
0.5	51.5	84.4	19.9
0.6	43.6	**85.7**	14.8
0.7	33.9	85.6	10.9
0.8	23.5	85.1	7.4
0.9	12.8	84.4	4.0

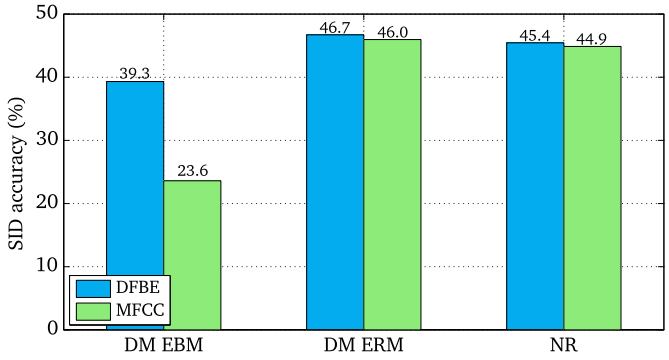

Fig. 4. SID performance of the DM approach with EBMs and ERMs in comparison to a conventional noise reduction stage ("NR") for both DFBE and FBE features. The performance is shown for 56 speakers at 0 dB SNR and averaged across all five noise types.

gap between the two DM systems employing either MFCC or DFBE features. When MFCCs were used ("MFCC DM"), each erroneously classified T-F unit affected all MFCC features due to the DCT, resulting in a distorted representation of the target signal. Considering the DFBE features ("DFBE DM"), the decorrelation was achieved by an FIR filter which only combined information from two neighboring frequency channels. Consequently, the distortions caused by misclassified T-F units remained local and did not affect a large number of DFBE features, which may explain the higher performance in comparison to the MFCC-based system. Evidently, there were large differences between the tested missing-data systems when comparing their performance under ideal and non-ideal conditions. In particular, the DM approach was very sensitive to estimation errors in the binary mask, which limited the SID accuracy in non-ideal conditions.

4.3. Experiment 3: DM versus noise reduction

To further analyze the sensitivity of the DM approach to errors in the binary mask, Fig. 4 compares the SID performance for both DFBE and FBE features using estimated binary masks and estimated ratio masks. Instead of binarizing the estimated SPP in the auditory domain $\widetilde{\mathcal{P}}(t, f)$ using (2), it was directly used as an estimation of the ratio mask. Similar to the binary masks, the same flooring value of 26 dB was applied to the ratio masks (see Section 2.3). In addition to the DM approach, the effect of a conventional noise reduction algorithm was tested by enhancing the STFT representation of noisy speech prior to feature extraction. Specifically, the minimum mean-square error (MMSE) gain function (Ephraim and Malah, 1984) was combined with the MMSE-based noise power estimation algorithm (Gerkmann and Hendriks, 2012), which was also used for the estimation of the SPP. Similar to the two previous experiments, a set of 56 speakers from the EMIME database was tested at 0 dB and results were averaged across 10 simulations and all five noise types.

It can be seen that the DM approach was much more effective when estimated ratio masks were used ("DM ERM") instead of estimated binary masks ("DM EBM"). Interestingly, there was almost no difference between the two feature representations when using either DM with ERMs or the MMSE-based noise reduction ("NR"), which highlights the sensitivity of the MFCC feature representation to errors in the binary mask. Moreover, despite the limited spectral resolution of 32 filters, the DM approach with ERMs performed slightly better than the noise reduction algorithm, which operated at a much higher spectral resolution in the STFT domain. In general, this experiment demonstrated the limitation of the DM approach when being used with binary masks. Although the ratio mask was substantially more effective than the binary mask,

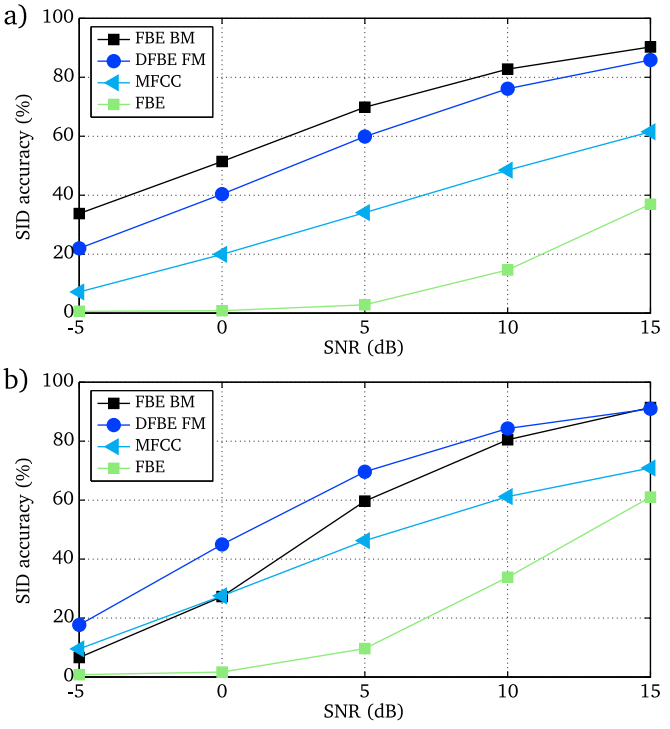

Fig. 5. SID performance for 200 speakers as a function of the SNR averaged across a) more stationary noises (ICRA1, 0.5-SAM and cockpit noise) and b) non-stationary noise (ICRA7 and factory noise).

the overall SID scores were almost 10% below the performance obtained by full and bounded marginalization reported in the previous experiment (see Section 4.2).

4.4. Experiment 4: comparison with MFCCs

Based on the second and third experiment, bounded marginalization with FBE features and full marginalization with DFBE features were found to be the most robust missing-data strategies under realistic conditions. Therefore, both approaches were compared with the frequently-used MFCC features over a wide range of SNRs using 200 speakers from the TIMIT database. Both missing-data systems employed their previously determined optimal SPP thresholds θ_{SPP}. The SID performance is presented in Fig. 5 as a function of the SNR for a) stationary and b) non-stationary background noises in separate panels. The corresponding evaluation of the EBM quality in terms of HIT - FA and mask accuracy is presented in Table 3.

Considering more stationary background noises (panel a), bounded marginalization ("FBE BM") was the most effective missing-data strategy which combined the information from both reliable and unreliable T-F units. As shown in Table 3, this was due the high accuracy of the estimated binary mask in stationary noises, which was above 90% at lower SNRs. However, the results were somewhat different for highly non-stationary and speech-modulated background noises as shown in panel b). Here, the SID performance of the BM approach ("FBE BM") decreased, because it was much more challenging to obtain an accurate estimation of the EBM. This is also reflected in Table 3, where the binary mask accuracy for non-stationary noises reduced by more than 15% at low SNRs compared to stationary noises. Instead of explicitly exploiting knowledge about unreliable T-F units, full marginalization simply ignores unreliable feature components, which apparently reduced the sensitivity to errors in the binary mask. As a result, FM ("DFBE FM") achieved the highest SID accuracy in the presence

Table 3
Evaluation of the EBMs produced by the two missing-data systems, "DFBE FM" and "FBE BM", as a function of the SNR and the stationarity of the background noise.

	Method	Metric (%)	SNR (dB)				
			−5	0	5	10	15
Stationary	DFBE FM	HIT - FA	59.3	65.1	69.3	72.1	74.0
		Accuracy	88.5	87.2	86.6	86.5	86.8
	FBE BM	HIT - FA	39.9	49.7	57.9	64.1	68.9
		Accuracy	92.7	90.0	87.8	86.0	84.8
Non-stationary	DFBE FM	HIT - FA	33.5	39.4	44.5	49.6	55.1
		Accuracy	67.3	69.2	71.7	75.0	78.7
	FBE BM	HIT - FA	26.3	34.8	42.0	48.9	56.1
		Accuracy	77.0	76.2	75.8	76.3	77.8

of non-stationary background noises. When comparing the MFCC-based system ("MFCC") to the FBE features with a conventional GMM classifier ("FBE"), a considerable improvement in SID accuracy was obtained. However, compared to the two missing-data approaches ("FBE BM" and "DFBE BM"), the MFCC features appeared to be less robust against the impact of background noise. Nevertheless, the ability of the two missing-data strategies, BM and FM, to deal with interfering noise was shown to strongly depend on the stationarity of the background noise.

4.5. Experiment 5: combining full and bounded marginalization

The previous experiment showed that it seemed advantageous to combine full marginalization and bounded marginalization due to their complementary performances in stationary and non-stationary background noises. Therefore, both systems were combined using a simple score fusion described in Section 2.7. In addition, an BM-based SID system with IBMs was used for comparison to indicate the upper performance limit of missing-data strategies. Figure 6 shows the SID performance for 200 speakers from the TIMIT database as a function of the SNR for a) stationary, b) non-stationary and c) all background noises in separate panels.

In general, the combined SID approach ("FBE BM & DFBE FM") was always better than the individual missing-data systems. In the presence of stationary noise (panel a), where the EBM was estimated with high accuracy, the SID performance of the combined system was dominated by the bounded marginalization approach ("FBE BM"). On the other hand, in case of non-stationary background noise (panel b), the combined system benefited from full marginalization ("DFBE FM"), which ignored unreliable T-F units, thereby decreasing the sensitivity to errors in the EBM. In comparison to the *ideal* SID system based on *a priori* knowledge ("FBE BM IBM"), the proposed combination obtained a fairly similar performance level in the presence of stationary noise (panel a) down to 5 dB SNR. Clearly, there is some room for improvements in non-stationary scenarios, due to the difficulty of obtaining an accurate estimation of the binary mask.

In summary, the high SID performance of the combined system can be attributed to the complementary error statistics of full and bounded marginalization and potential synergy effects between the different feature sets (FBE and DFBE features). A substantial performance benefit over an MFCC-based system ("MFCC") was achieved, without assuming any *a priori* knowledge about the interfering noise.

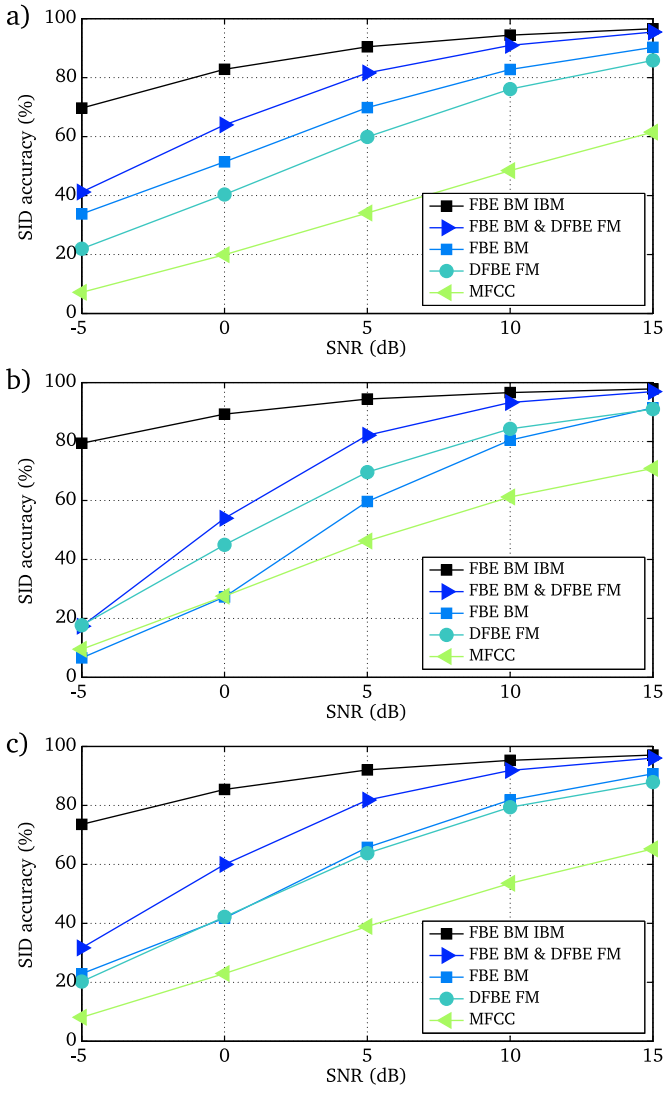

Fig. 6. SID performance for 200 speakers as a function of the SNR averaged across a) more stationary noises (ICRA1, 0.5-SAM and cockpit noise), b) non-stationary noise (ICRA7 and factory noise) and c) all five background noises.

5. Discussion and conclusion

This study compared the effectiveness of three missing-data strategies in the context of closed-set speaker identification, namely full marginalization (FM), bounded marginalization (BM) and direct masking (DM). A systematic evaluation under ideal and non-ideal conditions demonstrated that the performance of the different strategies was strongly affected by estimation errors in the binary mask.

Although BM and DM performed similarly under ideal conditions, the effectiveness of DM was drastically reduced when estimated binary masks were used. In particular, the combination of DM with MFCC features produced low SID scores under realistic conditions, presumably because each erroneously attenu-

ated speech-dominated T-F unit affected several MFC coefficients due the DCT stage, thereby distorting the resulting feature representation. In contrast, the FM approach produced similar SID scores compared to BM in conditions where the binary mask was estimated, despite showing the lowest potential of all missing-data strategies under ideal conditions. Interestingly, both FM and BM showed distinct advantages depending on the stationarity of the interfering noise. Whereas BM achieved the highest SID performance in the presence of stationary noise, the FM approach was more robust to non-stationary background noises. Apparently, there is a trade-off between the ability of the BM approach to benefit from the information contained in the unreliable T-F units and the accuracy of the estimated binary mask. Because FM does not explicitly utilize information from unreliable feature components, it is less sensitive to errors in the EBM, which are more likely to occur in non-stationary environments.

The combination of two missing-data systems, namely DM and BM, has been proposed in Zhao et al. (2014). However, the first three experiments in this study suggested that the DM strategy is particularly sensitive to estimation errors in the binary mask. Instead, the score fusion between FM and BM proposed here was shown to combine the distinct advantages of both missing-data strategies in stationary and non-stationary conditions.

Under non-ideal conditions, the EBM was derived by applying a threshold to the estimated speech presence probability in individual T-F units. It was demonstrated that optimal thresholds varied across missing-data strategies, reflecting their different requirements on the EBM. Whereas the optimal SPP threshold for both DM and FM maximized the HIT - FA rate, the BM approach achieved the highest SID performance when both speech-dominated and noise-dominated T-F units were correctly classified with similar priority, which was reflected by a high classification accuracy of the EBM. This is an important observation that must be considered when comparing binary mask estimators with different missing-data strategies. Often, algorithms that estimate the EBM have been optimized for a particular missing-data system, which complicates a transparent comparison. To assist the development of EBM estimators and to optimize their respective parameters, the two technical measures, namely the HIT - FA and the binary mask accuracy, seemed appropriate. Apart from evaluating the performance of binary masks on the basis of individual T-F units, the clustering of binary mask errors has been recently shown to strongly affect speech intelligibility (Kressner and Rozell, 2015), and its influence on SID performance should therefore be considered in future investigations.

The mask density, which reflects the percentage of speech-dominated T-F units in the EBM, can be quite low, thereby limiting the amount of information that is available for the identification task. In addition, the binary mask represents a binary decision, without reflecting the uncertainty of the underlying estimation. Among all tested strategies, the DM approach was shown to be particularly sensitive to estimation errors in the binary mask. To alleviate this, the binary decision about reliability can be softened by replacing the binary mask with the probability that individual T-F units are reliable (Barker et al., 2000). Such a ratio mask was shown to be superior to binary masks when being used in conjunction with the DM approach and its effectiveness for the classification-based missing-data strategies will be subject of further investigations.

In this study, the EBM was estimated without assuming any prior knowledge about the interfering background noise. Moreover, speaker models were trained with clean and anechoic speech, resulting in a very flexible SID system which does not require any prior knowledge about the acoustic environment. However, if concrete assumptions about the acoustic environment can be made, for example about the type of interfering noise, then the estimation of the binary mask can be treated as a binary classification problem. This would allow the use of supervised learning approaches, for instance based on amplitude modulation spectrogram features (May and Dau, 2014). These approaches were shown to substantially improve the accuracy of the EBM, at the cost of reducing the flexibility of the SID system.

Acknowledgments

This work has been supported by the European Union FP7 project TWO!EARS (http://www.twoears.eu) under grant agreement No. 618075.

References

Anzalone, M.C., Calandruccio, L., Doherty, K.A., Carney, L.H., 2006. Determination of the potential benefit of time-frequency gain manipulation. Ear Hear. 27 (5), 480–492. doi:10.1097/01.aud.0000233891.86809.df.

Barker, J., 2012. Missing-data techniques: Recognition with incomplete spectrograms. In: Virtanen, T., Singh, R., Raj, B. (Eds.), Techniques for noise robustness in automatic speech recognition. John Wiley & Sons, Ltd, New York, NY, USA. chapter 14, pp. 369–398. doi:10.1002/9781118392683.ch14.

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP, pp. 373–376.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. ICASSP, 4, pp. 208–211. doi:10.1109/ICASSP.1979.1170788.

Campbell, J.P., 1997. Speaker recognition: a tutorial. IEEE Trans. Audio, Speech, Lang. Process. 85 (9), 1437–1462. doi:10.1109/5.628714.

Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.-F., Matrouf, D., 2009. Forensic speaker recognition. IEEE Signal Process. Mag. 26 (2), 95–103. doi:10.1109/MSP.2008.931100.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34, 267–285. doi:10.1016/S0167-6393(00)00034-0.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Audio, Speech, Signal Process. 28 (4), 357–366. doi:10.1109/TASSP.1980.1163420.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech, Lang. Process. 19 (4), 788–798.

Deng, L., Droppo, J., Acero, A., 2005. Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion. IEEE Trans. Audio, Speech, Lang. Process. 13 (3), 412–421. doi:10.1109/TSA.2005.845814.

Dreschler, W.A., Verschuure, H., Ludvigsen, C., Westermann, S., 2001. ICRA Noises: artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. Audiology 40 (3), 148–157. doi:10.3109/00206090109073110.

Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environment with combined spectral subtraction and missing data theory. In: Proc. ICASSP, pp. 121–124. doi:10.1109/ICASSP.1998.674382.

El-Solh, A., Cuhadar, A., Goubran, R.A., 2007. Evaluation of speech enhancement techniques for speaker identification in noisy environments. In: Proc. ISMW, pp. 237–239. doi:10.1109/ISM.Workshops.2007.47.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Audio, Speech, Signal Process. 32 (6), 1109–1121. doi:10.1109/TASSP.1984.1164453.

Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V., 1993. TIMIT Acoustic-phonetic Continuous Speech Corpus. Technical Report NISTIR 4930. National Institute of Standards and Technology, Gaithersburg, MD, USA.

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio, Speech, Lang. Process. 20 (4), 1383–1393. doi:10.1109/TASL.2011.2180896.

Godin, K.W., Sadjadi, S.O., Hansen, J.H.L., 2013. Impact of noise reduction and spectrum estimation on noise robust speaker identification. In: Proc. Interspeech, pp. 3656–3660.

Hartmann, W., Narayanan, A., Fosler-Lussier, E., Wang, D.L., 2013. A direct masking approach to robust ASR. IEEE Trans. Audio, Speech, Lang. Process. 21 (10), 1993–2005. doi:10.1109/TASL.2013.2263802.

Jančovič, P., Köküer, M., 2006. Employment of voicing information of speech spectra for noise-robust speaker identification. In: Proc. EUSIPCO, pp. 2399–2403.

Jensen, J., Hendriks, R.C., 2012. Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions. IEEE Trans. Audio, Speech, Lang. Process. 20 (1), 92–102. doi:10.1109/TASL.2011.2157685.

Kim, G., Lu, Y., Hu, Y., Loizou, P.C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Amer. 126 (3), 1486–1494. doi:10.1121/1.3184603.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: from features to supervectors. Speech Commun. 52 (1), 12–40. doi:10.1016/j.specom.2009.08.009.

Kolossa, D., Haeb-Umbach, R., 2011. Robust speech recognition of uncertain or missing data. Springer, Berlin-Heidelberg, Germany doi:10.1007/978-3-642-21317-5.

Kressner, A.A., Rozell, C.J., 2015. Structure in time-frequency binary masking errors and its impact on speech intelligibility. J. Acoust. Soc. Amer. 137 (4), 2025–2035. doi:10.1121/1.4916271.

May, T., 2016. A MATLAB framework for missing data-based speaker identification experiments. Software is available at https://bitbucket.org/hea-dtu/speaker-identification-framework.

May, T., Dau, T., 2013. Environment-aware ideal binary mask estimation using monaural cues. In: Proc. WASPAA,New Paltz, NY, USA, pp. 1–4. doi:10.1109/WASPAA.2013.6701821.

May, T., Dau, T., 2014. Computational speech segregation based on an auditory-inspired modulation analysis. J. Acoust. Soc. Amer. 136 (6), 3350–3359. doi:10.1121/1.4901711.

May, T., Gerkmann, T., 2014. Generalization of supervised learning for binary mask estimation. In: Proc. IWAENC, pp. 154–158. doi:10.1109/IWAENC.2014.6953357.

May, T., van de Par, S., Kohlrausch, A., 2012a. A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. IEEE Trans. Audio, Speech, Lang. Process. 20 (7), 2016–2030. doi:10.1109/TASL.2012.2193391.

May, T., van de Par, S., Kohlrausch, A., 2012b. Noise-robust speaker recognition combining missing data techniques and universal background modeling. IEEE Trans. Audio, Speech, Lang. Process. 20 (1), 108–121. doi:10.1109/TASL.2011.2158309.

Nabney, I. T., Bishop, C. M., 2004. NETLAB package. Software is available at http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/. Last viewed July 2015.

Nadeu, C., Hernando, J., Gorricho, M., 1995. On the decorrelation of filter-bank energies in speech recognition. In: Proc. Eurospeech, pp. 1381–1384.

Nadeu, C., Macho, D., Hernando, J., 2001. Time and frequency filtering of filter-bank energies for robust HMM speech recognition. Speech Commun. 34 (1–2), 93–114. doi:10.1016/S0167-6393(00)00048-0.

Ozerov, A., Lagrange, M., Vincent, E., 2013. Uncertainty-based learning of acoustic models from noisy data. Comput. Speech Lang. 27 (3), 874–894. doi:10.1016/j.csl.2012.07.002.

Renevey, P., Drygajlo, A., 2000. Statistical estimation of unreliable features for robust speech recognition. In: Proc. ICASSP, vol. 3, pp. 1731–1734. doi:10.1109/ICASSP.2000.862086.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process. 10, 19–41. doi:10.1006/dspr.1999.0361.

Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. Speech Audio Process. 3 (1), 72–83. doi:10.1109/89.365379.

Ris, C., Dupont, S., 2001. Assessing local noise level estimation methods: application to noise robust ASR. Speech Commun. 34, 141–158. doi:10.1016/S0167-6393(00)00051-0.

Sadjadi, S.O., Hansen, J.H.L., 2010. Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. In: Proc. Interspeech, pp. 2138–2141.

Schädler, M.R., Kollmeier, B., 2015. Separable spectro-temporal gabor filter bank features: reducing the complexity of robust features for automatic speech recognition. J. Acoust. Soc. Amer. 137 (4), 2047–2059. doi:10.1121/1.4916618.

Seltzer, M.L., Raj, B., Stern, R.M., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Commun. 43 (4), 379–393. doi:10.1016/j.specom.2004.03.006.

Shao, Y., Srinivasan, S., Wang, D.L., 2007. Incorporating auditory feature uncertainties in robust speaker identification. In: Proc. ICASSP, pp. 277–280. doi:10.1109/ICASSP.2007.366903.

Shao, Y., Wang, D.L., 2006. Robust speaker recognition using binary time-frequency masks. In: Proc. ICASSP, pp. 645–648. doi:10.1109/ICASSP.2006.1660103.

Togneri, R., Pullella, D., 2011. An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag. 11 (2), 23–61.

de la Torre, A., Peinado, A.M., Segura, J.C., Perez-Cordoba, J.L., Benitez, M.C., Rubio, A.J., 2005. Histogram equalization of speech representation for robust speech recognition. IEEE Trans. Audio, Speech, Lang. Process. 13 (3), 355–366. doi:10.1109/TSA.2005.845805.

Varga, A.P., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251. doi:10.1016/0167-6393(93)90095-3.

Viikki, O., Laurila, K., 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Commun. 25 (1–3), 133–147. doi:10.1016/S0167-6393(98)00033-8.

Vizinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study. In: Proc. Eurospeech, pp. 2407–2410.

Wang, D.L., 2005. On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Kluwer Academic, Dordrecht, The Netherlands. chapter 12, pp. 181–197. doi:10.1007/0-387-22794-6_12.

Wester, M., 2010. The EMIME Bilingual Database. Tech. Rep. EDI-INF-RR-1388. University of Edinburgh.

Yu, C., Liu, G., Hahm, S., Hansen, J.H.L., 2014. Uncertainty propagation in front end factor analysis for noise robust speaker recognition. In: Proc. ICASSP, pp. 4017–4021. doi:10.1109/ICASSP.2014.6854356.

Zhao, X., Shao, Y., Wang, D.L., 2012. CASA-based robust speaker identification. IEEE Trans. Audio, Speech, Lang. Process. 20 (5), 1608–1616. doi:10.1109/TASL.2012.2186803.

Zhao, X., Wang, Y., Wang, D.L., 2014. Robust speaker identification in noisy and reverberant conditions. IEEE/ACM Trans. Audio, Speech, Lang. Process. 22 (4), 836–845. doi:10.1109/TASLP.2014.2308398.